Analysis of ongoing neuronal activity evoked by continuous speech with computational corpus linguistics methods

Achim Schilling[1,2], Rosario Tomasello[3,4], Malte R. Henningsen-Schomers[3,4], Kishore Surendra[1,2,5], Martin Haller[1,2], Valerie Karl[6], Peter Uhrig[7,8], Andreas Maier[5], and Patrick Krauss[1,2,8,9]

[1]Cognitive Computational Neuroscience Group at the Chair of English Philology and Linguistics, Friedrich-Alexander University Erlangen-Nürnberg (FAU), Germany
[2]Neuroscience Lab, Experimental Otolaryngology, University Hospital Erlangen, Germany
[3]Brain Language Laboratory, Department of Philosophy and Humanities, Freie Universität Berlin, Berlin, Germany
[4]Cluster of Excellence "Matters of Activity. Image Space Material", Humboldt Universität zu Berlin, Germany
[5]Chair of Machine Intelligence, Friedrich-Alexander University Erlangen-Nürnberg (FAU), Germany
[6]Biomagnetism Lab, Department of Neurosurgery, University Hospital Erlangen, Germany
[7]Chair of English Philology and Linguistics, Friedrich-Alexander University Erlangen-Nürnberg (FAU), Germany
[8]FAU Linguistics Lab, Friedrich-Alexander University Erlangen-Nürnberg (FAU), Germany
[9]Department of Otorhinolaryngology/Head and Neck Surgery, University of Groningen, University Medical Center Groningen, The Netherlands

June 2, 2020

Corresponding author:
Dr. Patrick Krauss
Cognitive Computational Neuroscience Group
at the Chair of English Philology and Linguistics
Friedrich-Alexander University of Erlangen-Nürnberg (FAU)
Glückstrasse 5
91054 Erlangen, Germany
patrick.krauss@fau.de

Keywords:
neurolinguistics, MEG/EEG, event related potentials, event related fields, natural language processing, forced alignment, naturalistic speech stimuli

Abstract

In the field of neurobiology of language, neuroimaging studies are generally based on stimulation paradigms consisting of at least two different conditions. Designing those paradigms can be very time-consuming and this traditional approach is necessarily data-limited. In contrast, in linguistics analyses are often based on large text corpora, which allow a vast variety of hypotheses to be tested by repeatedly re-evaluating the data set. Furthermore, text corpora also allow exploratory data analysis in order to generate new hypotheses. By combining the two approaches, we here present a unified approach of continuous natural speech and MEG to generate a corpus-like database of speech-evoked neuronal activity.

Introduction

Contemporary linguistic research is characterized by a great variety of methodological approaches. The fields of psycholinguistics and neurobiology of language in particular are characterized by a vast number of different methods that are applied in order to investigate the neuronal and cognitive correlates and processing principles of language acquisition, representation, comprehension and production [1]. In neurolinguistics, electrophysiological measurements, i.e. magnetoencephalography (MEG) [2] and electroencephalography (EEG) [3, 4], are widely used to investigate the neural correlates of language processing in the human brain [5–10].

However, most of the experimental studies on language processing conducted so far have focused on one aspect of linguistic information at a time. For instance, neurocognitive studies have explored the neural responses of words compared to pseudo words [11, 12], between different conceptual semantic categories [13], complex against simple grammatical sentences [14], or during pragmatic processing of different communicative actions [10]. Although, all these studies shed light on the correlates of language processing in the human brain, it is still not fully understood whether similar brain responses during single words or sentence understanding also emerge during perception of natural speech, similar to everyday experience. However, recently a growing number of approaches address this issue [15–20].

Furthermore, traditional experimental designs typically consist of at least two different conditions studied under carefully controlled circumstances [7–9]. The measured data are then pre-processed, i.e. referenced, filtered, epoched and averaged, and finally contrasted according to the different stimulation conditions [1]. To obtain a good signal to noise ratio (SNR) of the acquired brain responses, each of these conditions must contain dozens of different items or stimulus repetitions.

For instance, the evaluation of event related potentials (ERPs) from the EEG data, or, in the case of MEG, event related fields (ERFs), requires a relatively large number of stimuli (40-120 trials) per condition to achieve high SNR and ensure sufficient statistical power. This is due to the fact that the signal of a specific condition remains constant across multiple repetitions while the noise signal which is assumed to be randomly distributed, is reduced when large number of time-locked stimuli are pooled together [21–25]. Apart from creating a large number of stimuli to increase SNR, it is also well known that neural activity is reduced by stimulus repetition, the so-called

repetition suppression [26]. Thus, it is necessary to design a certain number of different stimuli from each condition to avoid repetition, which is often very challenging or even impossible. One strategy is to focus on single-item ERPs/ERFs, but in such cases it is necessary to compensate by testing more participants to obtain stable signals [27].

Here, we present an alternative approach to overcome the aforementioned limitations of the electrophysiological assessment of language processing, and to open up the possibility of investigating different levels of linguistic information during natural speech comprehension within a single experiment. In particular, in the present study, we investigated brain responses elicited during listening to the audiobook edition of a German-language novel by means of MEG measurements (for similar approaches see [28, 29]). Using computational corpus linguistics (CCL) [30, 31] applied to the analyses of large text corpora, which usually consist of hundreds of thousands or even billions of tokens [32–37], offers the opportunity to test a vast number of hypotheses by repeatedly re-analyzing the data [38] and to deploy modern machine learning techniques on such datasets [39]. Furthermore, text corpora also allow for exploratory data analyses in order to generate new hypotheses [40]. In our approach we can generate a large database of neuronal activity in a single measurement session, corresponding to the comprehension of several thousands of words of continuous speech similarly to everyday language. However, for later studies the data set has to be split into multiple parts (e.g. development/training/test or training/validation/test) in analogy to standard machine learning data sets as MNIST (50,000 training images, 10,000 test images [41]), as hypothesis generation and checking for statistical significance have to be done in two disjoint steps, in order to prevent *harking* (hypothesizing after the results are known) [42]. In such cases, inferential statistical analysis is not valid and applicable [43]. Thus, this approach is only possible with a large dataset.

Here, we provide the proof-of-principle of this approach by calculating ERFs and normalized power spectra of word onsets and offsets overall as well as for the group of content words (nouns, verbs, adjectives) and for the group of function words (determiners, prepositions, conjunctions), which are known to differ semantically to a substantial extent. Hence, greater activation for content compared to function words can be expected, as reported in previous studies (e.g., [44, 45]) [1]. Furthermore, we check for consistency

[1] We follow the categorization of these studies and thus only include the content and

of the data, by comparing intra-individual differences of neural activity in different brain regions and we perform non-parametric cluster permutation tests to determine significant differences between conditions.

function word classes listed above in our study.

Methods

Human participants

Participants were 15 (8 females, 7 males) healthy right-handed (augmented laterality index: $\mu = 85.7$, $\sigma = 10.4$) and monolingual native speakers of German aged 20-42 years. They had normal hearing and did not report any history of neurological illness or drug abuse. They were paid for their participation after signing an informed consent form. Ethical permission for the study was granted by the ethics board of the University Hospital Erlangen (registration no. 161-18 B). For the questionnaire based assessment and analysis of handedness we used the Edinburgh Inventory [46].

Speech stimuli and natural language text data

As natural language text data, we used the German novel *Gut gegen Nordwind* by Daniel Glattauer (© *Deuticke im Paul Zsolnay Verlag*, Wien 2006) which was published by *Deuticke Verlag*. As speech stimuli, we used the corresponding audiobook which was published by *Hörbuch Hamburg*. Both, the novel and the audiobook are available in stores, and the respective publishers gave us permission to use them for the present and future scientific studies.

Book and audiobook consist of a total number of 40460 tokens (number of words) and 6117 types (number of different words). The distribution of uni-gram word classes and bi-gram word class combinations occurring in the text were analysed by applying *part-of-speech (POS) tagging* [47, 48] as implemented in the python library *SpaCy* [49]. The results are summarized in Figure 1.

The total duration of the audiobook is approximately 4.5 hours. For our study, we only used the first 40 minutes of the audiobook, (which corresponds to approximately 6000 words of spoken language) divided into 10 parts of approximately 4 minutes ($\mu = 245\,s$, $\sigma = 39\,s$). In order to avoid cutting the text in the middle of a sentence or even in the middle of a word, we manually cut at paragraph boundaries, which resulted in more meaningful interruptions of the text. For the present study, only the first 3 sections (roughly 12 minutes) of the recordings and the corresponding measurements were analyzed.

Stimulation protocol

The continuous speech from the audiobook was presented in 10 subsequent parts (cf. above) at a sensory level of approximately 30-60 dB SL. The actual loudness varied from participant to participant. It was chosen individually to ensure good intelligibility during the entire measurement, but also to prevent it from being unpleasant. Simultaneously with auditory stimulation, a fixation cross at the centre of the screen was presented all the time to minimize artifacts from eye movements. After each audiobook part, three multiple-choice questions on the content of the previously presented part were presented in order to test the participants' attention. Furthermore, stimulation was interrupted for a short break of approximately 5 minutes after audiobook parts number 4 and 7. The total duration of the protocol is approximately one hour. The complete stimulation protocol is shown in Figure 2.

Generation of trigger pulses with forced alignment

In order to automatically create trigger pulses for both, the synchronization of the speech stream with the MEG recordings, and to mark the boundaries of words, phonemes, and silence for further segmentation of the continuous data streams, *forced alignment* [50–52] was applied to the text and recording. For this study we used the free web service *WebMAUS* [53, 54]. It takes a wave file containing the speech signal, and a corresponding text file as input and gives three files as output: the time tags of word boundaries, a phonetic transcription of the text file, and the time tags of phone boundaries. Even though forced alignment is a fast and reliable method for the automatic phonetic transcription of continuous speech, we carried out random manual inspections in order to ensure that the method actually worked correctly. For simplicity, we only used the time tags of word boundaries in this study. However, a more fine grained analysis on the level of speech sounds could easily be performed retrospectively, since the time tags of beginning and ending of a given word correspond to the beginning of the word's first phone and the ending of the word's last phone, respectively. Thus, the two lists containing the time tags of the word and phoneme boundaries can easily be aligned with each other. Although forced alignment is not 100% reliable, manual spot checks found no errors in our alignment. Of course, the high-quality recording of an audiobook is among the best possible inputs for such

software.

Setup for speech presentation and synchronisation with the MEG recordings

The speech signal was presented using a custom made setup (Figure 3). It consists of a stimulation computer connected to an external USB sound device (Asus Xonar MKII, 7.1 channels) providing five analog outputs. The first and second analog outputs are connected to an audio amplifier (AIWA, XA-003), where the first output is connected in parallel to an analog input channel of the MEG data logger in order to enable an exact alignment of the presented stimuli and the recorded MEG signals (cf. Figure 3 a). In addition, the third analog output of the sound device is used to feed the trigger pulses derived from forced alignment into the MEG recording system via another analog input channel. In doing so, our setup prevents temporal jittering of the presented signal caused by multi-threading of the stimulation PC's operating system, for instance. For an overview of the wiring scheme of all devices see Figure 3a.

The speech sound was transmitted into the magnetically shielded MEG chamber to the participants' ears via a custom-made device consisting of two loudspeakers (Pioneer, TS-G1020F) which are coupled to silicone funnels each connected to a flexible tube of $\approx 2\,m$ length and with an inner diameter of $\approx 2\,cm$ (Figure 3b). These tubes are led through a small hole in the magnetically shielded chamber to prevent artifacts produced by interfering magnetic fields generated by the loudspeakers (Figure 3c). We carried out calibration tests to ensure that the acoustical distortions caused by the tube system do not affect speech intelligibility. Furthermore, due to the length of the tubes and the speed of sound, there is a constant time delay from the generation of sound to the arrival of the sound at the participant of $\approx 6\,ms$, which we took into account for the alignment described below.

The stimulation software is implemented using the programming language *Python 3.6*, together with Python's sound device library, the *PsychoPy* library [55, 56] for the stimulation protocol, and the *NumPy* library for basic mathematical and numerical operations.

Magnetoencephalography and data processing

MEG data (248 magnetometers, 4D Neuroimaging, San Diego, CA, USA) were recorded (1017.25 Hz sampling rate, filtering: 0.1-200 Hz analogue band pass, supine position, eyes open) during speech stimulation. Positions of five landmarks (nasion, LPA, RPA, Cz, inion) were acquired using an integrated digitizer (Polhemus, Colchester, Vermont, Canada). MEG data were corrected for environmental noise using a calibrated linear weighting of 23 reference sensors (manufacturers algorithm, 4D Neuroimaging, San Diego, CA, USA).

Further processing was performed using the Python library *MNE* [57,58]. Data were digitally filtered offline (1-10 Hz bandpass for ERF analyses; 50 Hz notch on for power spectra analysis) and downsampled to a sampling rate of 1000 Hz. MEG sensor positions were co-registered to the ICBM-152 standard head model [59] and atlas 19-21 [60], as individual MRI data sets for the participants were not available. Furthermore, recordings were corrected for eye blinks and electrocardiography artifacts based on signal space projection of averaged artifact patterns, as implemented in *MNE* [57, 58].

Additionally, we performed an independent component analysis (ICA) and deleted the first two independent components of the data, to further improve data quality. However, it appears that this processing step does not affect the observed differences of neural responses to function and content words (Figures 7 and S10).

Trials with amplitudes higher than $2 \cdot 10^{-12}$ T would have been rejected, as they were supposed to arise from artifacts. However, none of the trials fit this condition, and hence no trail was rejected.

For this study, we restricted our analyses to sensor space, and did not perform source localization in analogy to other ERF studies [61–63].

Alignment and segmentation

Since we have both, the original audiobook wave file together with the time tags of word boundaries from forced alignment (Figure 4a), and the corresponding recordings of two analog auxiliary channels of the MEG (Figure 4b), all 248 MEG recording channels could easily be aligned offline with the speech stream (Figure 4c). Subsequently, the continuous multi-channel MEG recordings were segmented using the time tags as boundaries, and labeled with the corresponding types, in our case individual words (Figure 5). For

the analysis of function and content words, we additionally applied POS tagging [47,48] using SpaCy [49] to assign word classes (nouns, verbs, adjectives, conjunctions, determiners, prepositions) to the individual words. Note that, in principle, the process of segmentation can also be performed at different levels of granularity. For instance, using the time tags of phone boundaries would result in a more fine-grained segmentation, whereas grouping several words together to n-grams with appropriate labels to larger linguistic units (i.e. collocations, phrases, clauses, sentences) would result in a more coarse-grained segmentation.

Event related fields

In order to provide the proof-of-principle of our approach, we analysed event related fields (ERF) evoked by word onsets (Figure 6). Since the continuous MEG signals of all 248 channels are already segmented according to word boundaries, we can compute ERFs of word onsets for each channel by simply averaging the pre-processed signals over the word tokens in our database. Here, we included only those words that follow a short pause, instead of using all words occurring in the data set. Thus, there is a short period of silence, ranging from approx. 50 ms to 1.5 seconds, before the actual word onsets which improves signal quality, yet with the drawback that only a fraction of all tokens can be used. However, there were still 291 remaining events, baseline corrected, within the first three parts of the audiobook, corresponding to approx. 12 min of continuous speech, that fit this condition.

In addition, we also analysed ERFs evoked by prototypical content words (nouns, verbs, adjectives) and compared them with ERFs evoked by function words (determiners, prepositions, conjunctions). Again, we included only those words that follow a short interval of silence, instead of using all words occurring in the data set. Within the first three parts of the audiobook, this resulted in 81 remaining events for the content word condition and 106 events for the function word condition. [2]

Permutation test

We performed intra-individual permutation tests to determine the p-value for the ERF comparison between content and function word. Thus, the ERF

[2] The numbers do not add up to the total of 291 words because only the word classes listed in the introduction were included in the analysis of content and function words.

was cut into four subsequent time frames, each with a duration of 250 ms, and the root-mean-square amplitude (RMS) was calculated (Figure 7e):

$RMS = \sqrt{\frac{1}{N}\sum_{i=0}^{N} v_i^2}$, with the signal values within a 250 ms interval v_i, the total number of values within a 250 ms interval $N = f_s \cdot 250\,\text{ms}$, and the sampling rate $f_s = 1000\,\text{Hz}$.

1000 different random permutations of content word and function word labels were generated. For each of these samples the four RMS amplitudes for the different time frames were calculated based on the baseline corrected single trials, resulting in a distribution of amplitudes for each time frame (Figure 7f). The amplitude values corresponding to the true labeling are compared with amplitudes derived from random permutations in order to estimate the statistical significance, i.e. the p-values for content and function words p_{con} and p_{fun}, respectively, in Figure 7f.

Normalized power spectra

Using Fourier transformation, we also analyzed the averaged normalized power spectra (alpha, beta and gamma frequency range) for words in contrast to pauses, and for function and content words. The frequency band were defined as follows : $\alpha : 8 - 12\,\text{Hz}$, $\beta : 12 - 30\,\text{Hz}$, $\gamma : 30 - 45\,\text{Hz}$. The epoch length for this analysis was 1 s for both, short periods of silence and word onsets. Furthermore, we projected the resulting values to the corresponding spatial position of the sensors. This was done by the usage of the plot_psd_topomap function of the *MNE* library with Python interface [57,58].

Results

The general idea of our approach was to perform MEG measurements of participants listening to an audiobook. By synchronizing the continuous speech stream with the ongoing multi-channel neuronal activity, and subsequently automatically segmenting the data streams according to word boundaries derived from forced alignment, we generated a database of annotated speech evoked neuronal activity. This corpus-like database may then be analyzed offline by applying the full range of methods from statistics, natural language processing, and computational corpus linguistics. In order to demonstrate the feasibility of our approach, we restricted our analyses to sensor space and did not perform any kind of source localization (cf. Methods). More specifically, we calculate averaged ERFs for word onsets, and normalized power spectra for onsets of both, words and short pauses. In addition, we compare averaged ERFs for content and function words, and the corresponding normalized power spectra and discuss the results in the light of existing studies.

Event related fields of word onsets

For a first proof of concept and to determine clear neurophysiological brain responses from continuous speech, we analyzed event related fields (ERFs) for word onsets (irrespective of their word classes) from different topographical sides.

Figure 6a shows one example of the resulting ERFs averaged over the aforementioned 291 events corresponding to word onsets for one participant (subject 2 of 15) and parts number 1 to 3 of the audiobook and Figure 6b shows a projection of the spatial distribution of the ERF amplitudes at 350 ms after word onset. The largest amplitudes occur in channels located at temporal and frontal areas of the left hemisphere known to be associated with language processing [64]. The ERFs of those channels with the largest ERF amplitudes are shown in Figure 6c. Furthermore, we see a clear N400 component for each condition, indicating language associated processing (cf. [17, 65–69]).

In order to exclude random effects, we compare these channels with the corresponding channels located at the right hemisphere – where we expect less activation due to the asymmetric laterilsation of speech in the brain – (Figure 6d), and with some occipital channels (Figure 6e). In both cases, the

resulting ERF amplitudes are clearly smaller than those of the left temporal and frontal channels (Figure 6c). In addition, we calculate control ERFs for the same channels shown in Figure 6c, but instead of word boundaries we used randomly chosen time tags for segmentation. Also in this control condition, the resulting ERF amplitudes are smaller than those for the word onset condition (Figure 6f). This result,in particular, demonstrates that even though there are no or only relatively short inter stimulus intervals, leading to overlapping effects of late and early responses of subsequent words, there is still enough signal left in the individual trials.

Finally, we evaluated the re-test reliability of our results using three-fold sub-sampling by separately averaging only over events belonging to the same part of the the audiobook (Supplemental Figures S1, S2, S3). Again, the largest ERF amplitudes were found in the same channels as before and all results show very similar patterns to those shown in Figure 6. In addition, we provide exemplary results of two further participants in the Supplements section (Supplemental Figures S4 and S5).

Event related fields of content and function words

As a further validity test of the present study, we analyzed and compared the brain responses of function words, defined as determiners, prepositions and conjunctions, and content words, defined as nouns, verbs and adjectives. We choose these lexical categories as they greatly differ in the semantic domain. As an example, the resulting ERFs averaged over the respective events (content words: n=81, function words: n=106) for one participant (subject 2 of 15) and parts number 1 to 3 of the audiobook are shown in Figure 7a,c. and a projection of the spatial distribution of the ERF amplitudes at 550 ms after word onset is provided in Figure 7b,d. We found that content words (Figure 7a,b) elicited greater activation than function words (Figure 7c,d), especially in temporal and frontal areas of the left hemisphere. Since content parts of speech have been shown to differ semantically from function parts of speech [70, 71], this finding is in line with previous studies [45]. In addition, we compared the two conditions for the channel yielding the largest ERF amplitude and performed a permutation test independently for four subsequent time frames each with a duration of 250 ms (Figure 7e,f). We found that the averaged ERFs for the two conditions (content and function words, intra-individual) are significantly ($p < 0.05$) different within the first (0 ms–250 ms) and third (500 ms–750 ms), but not within the second (250 ms–500 ms) and

fourth (750 ms–1000 ms) time frame (Figure 7f, two further subjects shown in Figures S6, S7).

Averaged normalized power spectra

In our analysis of the averaged normalized power spectra we were unable to find significant differences between the conditions of word onset and of silence onset (Supplemental Figure S8), and neither between content words and function words (Supplemental Figure S9). See discussion section for possible reasons.

Discussion

In this study, we presented an approach where we combine electrophysiological assessment of neuronal activity with computational corpus linguistics, in order to create a corpus-like database of continuous speech-evoked neuronal activity. We demonstrated that using an audiobook as natural speech stimulus, and simultaneously performing MEG measurements led to a relatively large number of analysable events (word onsets: $n = 291$, silence onsets: $n = 187$, content words: $n = 81$, function words: $n = 106$), yet within a relatively short measurement time of $15\,min$. We further provided the proof-of-principle that, in contrast to common study designs, even though our stimulus trials were not presented in isolation, i.e. with appropriate inter-stimulus intervals of a few seconds, averaging over all respective events of a certain condition results in ERFs in left temporal and frontal channels with increased amplitudes compared to those of several control channels (e.g. at right hemisphere or at occipital lobe). The same is true with respect to comparison with control conditions (e.g. random trigger times). These results are well in line with previously published findings [64]. Furthermore, we analyzed ERFs for content words and function words. We found a clear N400 component (cf. Figure 7) especially in left hemispheric frontal regions for both function and content words and a positive component from 400-700 ms which is in line with Brennan et al.'s findings [72, 73].

Additionally, we found that content words elicit greater activation than function words, especially in temporal and frontal areas of the left hemisphere. Due to their substantial semantic differences [70, 71], this finding is in line with previous studies [45].

With respect to the average normalized power spectra, it was found that presentation of speech stimuli was associated with an increase in broadband gamma and a decrease in alpha over auditory cortex, while alpha power was increased in domain unspecific cortical areas [74–76]. One reason could be that, since we analysed only very short periods of silence, i.e. between two words, our two conditions of word onset and silence onset can be considered basically, at a larger time scale, to be the same condition, i.e. continuous speech stimulation. This may explain why we found no differences in frequency power here. Even though it has been proposed that in human language networks linguistic information of different types is transferred in different oscillatory bands – in particular attention is assumed to correlate with an increase in gamma and a decrease in alpha band power [77] –, the

role of different spectral bands in mediating cognitive processes is still not fully understood. Therefore it remains unclear, whether these findings extend to content and function words. Whether our approach is too insensitive to see differences here remains to be seen, and further studies should look more closely at this issue.

As mentioned above, in contrast to traditional studies that are limited to testing only a small number of stimuli or word categories, the present approach opens the possibility to explore the neuronal correlates underlying different word meaning information across a large range of semantic categories [29], and syntactic structures [78]. This is because the ongoing natural speech used here contains both, a large number of words from different semantic domains [28] and a large number of sentences at all levels of linguistic complexity [79].

On the other hand, one may argue that stimulation with ongoing natural speech has, compared to traditional approaches, the drawback that there are virtually no inter stimulus intervals between the single words. This, of course, introduces a mixture of effects at different temporal scales, e.g. early responses to the actual word are confounded with late responses of the previous word. However, all these effects may be averaged out, as demonstrated by other studies [15–20] and also by our results.

In a follow-up study, it will have to be validated whether our approach also works for linguistic units of different complexity other than single words. For instance, smaller linguistic units like phonemes and morphemes, but also larger linguistic units like collocations, phrases, clauses, sentences, or even beyond, could be investigated. For instance, we might be able to determine what neural correlates of the different association measures used in research on collocation look like (see [80] for an overview and further references). Furthermore, more abstract linguistic phenomena need to be analyzed, e.g. argument structure constructions [81–83] or valency [84–86]. Finally, our speech-evoked neural data may also be grouped, averaged, and subsequently contrasted according to male and female voice, looking at gender specific differences (see e.g., [87, 88]).

Also, analyses based on source space need to be tested, as well as more sophisticated analyses taking advantage of the multi-dimensionality of the data, such as, for instance, multi-dimensional cluster statistics [89] and state-of-the-art deep learning approaches for creating so-called embeddings of the raw data [90]. Due to the corpus-like features of our data, all additional analyses mentioned may be performed on the existing database, and without

the need for designing new stimulation paradigms, or carrying out additional measurements.

However, in order to avoid statistical errors due to *HARKing* [42, 43] - defined as generating scientific statements exclusively based on the analysis of huge data sets without previous hypotheses - and to guarantee consistency of the data, it is necessary to apply e.g. re-sampling techniques such as sub-sampling as shown above and described in detail in [91]. Furthermore, the approach presented here allows us to apply the well-established machine learning practice of dataset splitting, i.e. to split the dataset into multiple parts before the beginning of the evaluation, where the one part is used for generating new hypotheses, and another part for subsequently testing these hypotheses (or split again into training and testing data).

To conclude, there are two major reasons why we think the study of the neurobiology of language can benefit tremendously from the introduction of corpus-linguistic methodology. The first is that we can base our research on naturally occurring language, which should make them more ecologically valid than the more artificial stimuli used in carefully balanced and controlled experiments. Of course, even though audiobooks are frequently used in similar studies [15–20], one may also discuss whether audiobooks actually can be considered natural speech. One could argue that the fact that highly trained professional speakers and actors are usually employed to read audiobooks, who may use specific intonational patterns to paint a more vivid image of the situation, may lead to unnaturalness and thus possibly to unusual arousal patterns in the hearer. However, this argument is flawed. People spend large portions of their days listening to language produced by such professional speakers for radio, television news and drama, online videos, and podcasts. While probably not predominant for most people, it corresponds to a perfectly normal, everyday type of language experience. Even if we expect deviations from spoken interaction in such stimuli, we could even exploit this to study brain responses to creative language use (see [92], [93] and the sources cited there for linguistic studies of creativity). Of course, further studies using recordings of everyday dialogues between untrained subjects, e.g. describing what they have done during the day, should be designed to obtain a more comprehensive picture and more robust results, because, as Kriegeskorte and Douglas pointed out: *"As we engage all aspects of the human mind, our tasks will need to simulate natural environments"* [94]. Still, purely receptive task such as the one used in this study is one type of natural environment, and one that can be studied without too much interference

compared to, say, spontaneous interaction.

The second reason is the fact that measurements can be re-used if they form part of a large corpus of neuroimaging results. Let us look at a few numbers: In the present study, we stimulated 15 participants with 40 minutes of audio each. Test time spent in the MEG was 60 minutes due to the questions and pauses mentioned above. With 30 minutes of preparation, we used the MEG lab for a total of 22.5 hours during experimentation. In that period of time, we gathered measurements for roughly 6,000 words perceived by 15 participants, totalling 90,000 sets of brain responses to words. These correspond to roughly 35 GB of measurements (4 bytes per value, 1,000 per second, 248 channels, 40 minutes per participant, 15 participants). For this study, we only looked at a tiny fraction of the data (words preceded by a short pause of at least 50 ms in the first 12 minutes) and already managed to confirm certain patterns found by previous studies with a strict experimental design. If we assume that pauses are equally distributed across the corpus, we can expect to find roughly 1,000 such events, with 15 participants for each, i.e. 15,000 datapoints alone for words preceded by silence. Having these plus all the other words in their immediate linguistics contexts without pauses opens many more avenues for interesting research question at no added laboratory costs. Once we start looking at all words, we expect that the noise introduced through not being able to control for a variety of factors will be counterbalanced by the sheer size of datasets constructed using the methodology presented.

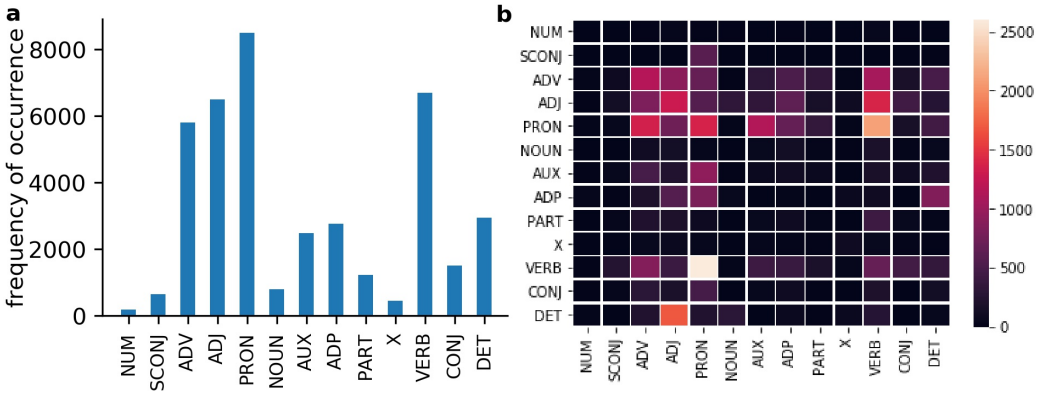

Figure 1: **Distribution of word classes and bi-gram word classes.** a: Distribution of word classes according to POS-tagging. Pronouns (PRON), verbs (VERB), adjectives (ADJ) and adverbs (ADV) are the most frequent word classes. The less frequent are: numbers (NUM), subordinating conjunctions (SCONJ), nouns (NOUN), auxiliary verbs (AUX), adpositions (ADP), particles (PART), conjunctions (CONJ), determiners (DET), and others (X). b: Distribution of word classes of 2-word sequences. The two most frequent sequence types are VERB-PRON and PRON-VERB. Rows: word class of first word. Columns: word class of second word.

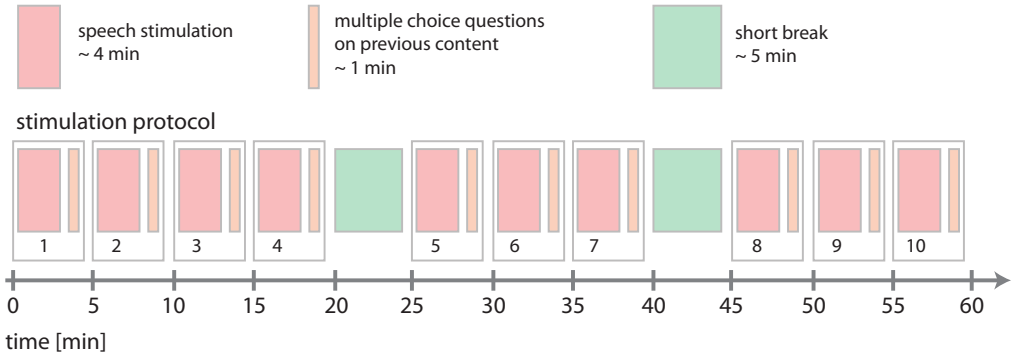

Figure 2: **Stimulation protocol.**
The total duration of the protocol was approximately one hour. The audiobook was presented in 10 subsequent parts with an average duration of 4 minutes (red). After each part, three multiple-choice questions on the content of the previous part of the audiobook were presented (orange). After audiobook parts number 4 and 7, stimulation was interrupted for a short break of approximately 5 minutes (green).

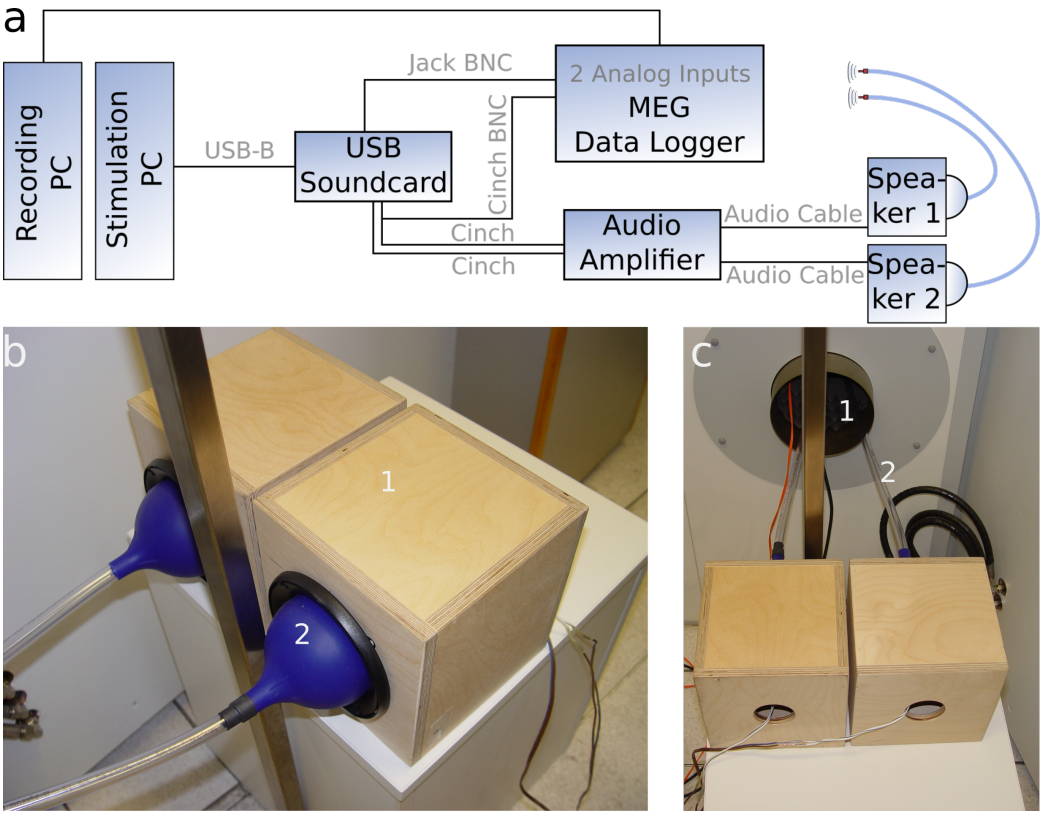

Figure 3: **Setup configuration.**
a: Wiring scheme of the different devices. b: The speech sound is transmitted into the magnetically shielded chamber via a custom-made construction consisting of two loudspeakers (1) which are coupled to silicone funnels (2) each connected to a flexible tube. c: Through a small whole in the magnetically shielded chamber (1), speech sound is transmitted via the two flexible tubes (2).

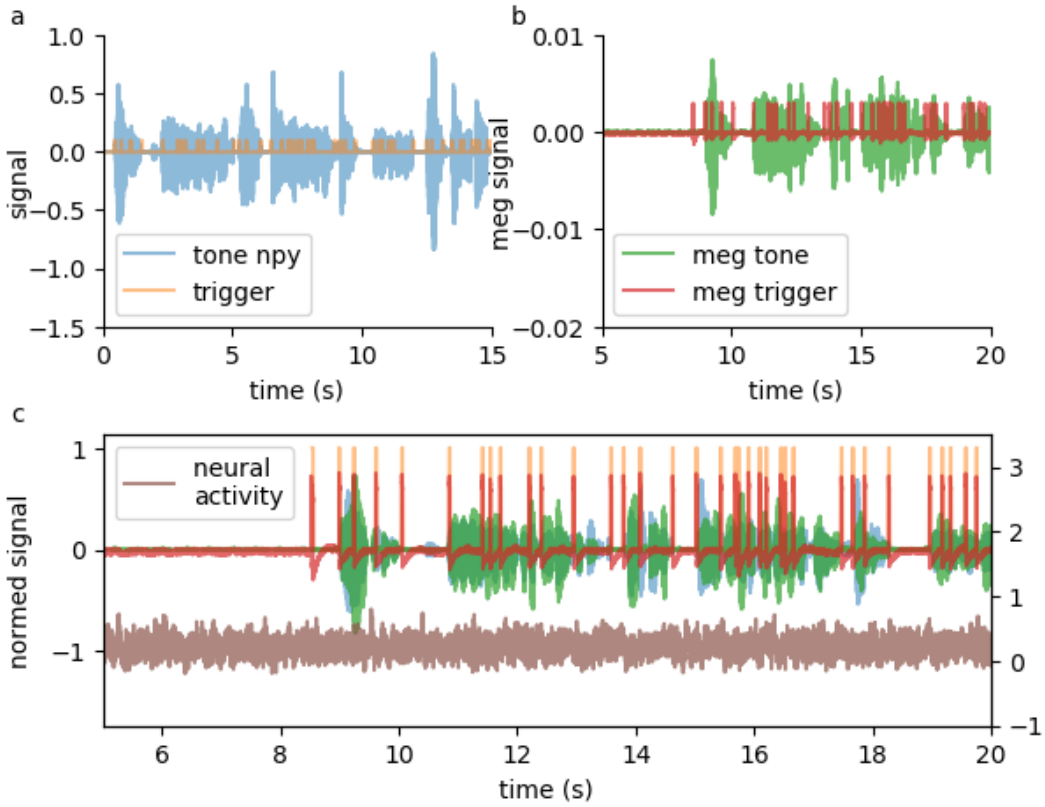

Figure 4: **Alignment of speech stream and MEG signal.**
a: Sample audiobook wave file (blue) together with time tags of word boundaries (orange) from forced alignment. b: Corresponding recordings of two analog auxiliary channels of the MEG. c: Alignment of data streams from a and b, together with one sample MEG channel.

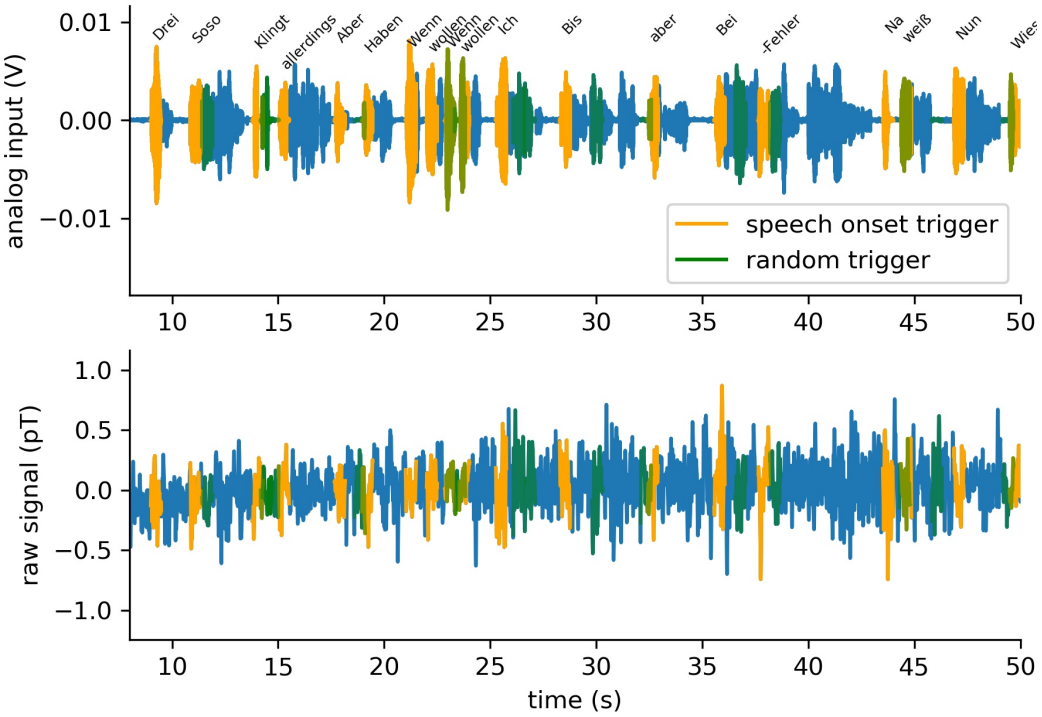

Figure 5: **Segmentation of speech stream and MEG signal.**
After alignment, the continuous wave file (top panel) and multi-channel MEG recordings (bottom panel) are segmented using the time tags from forced alignment as boundaries, and labeled with the corresponding types, i.e. words.

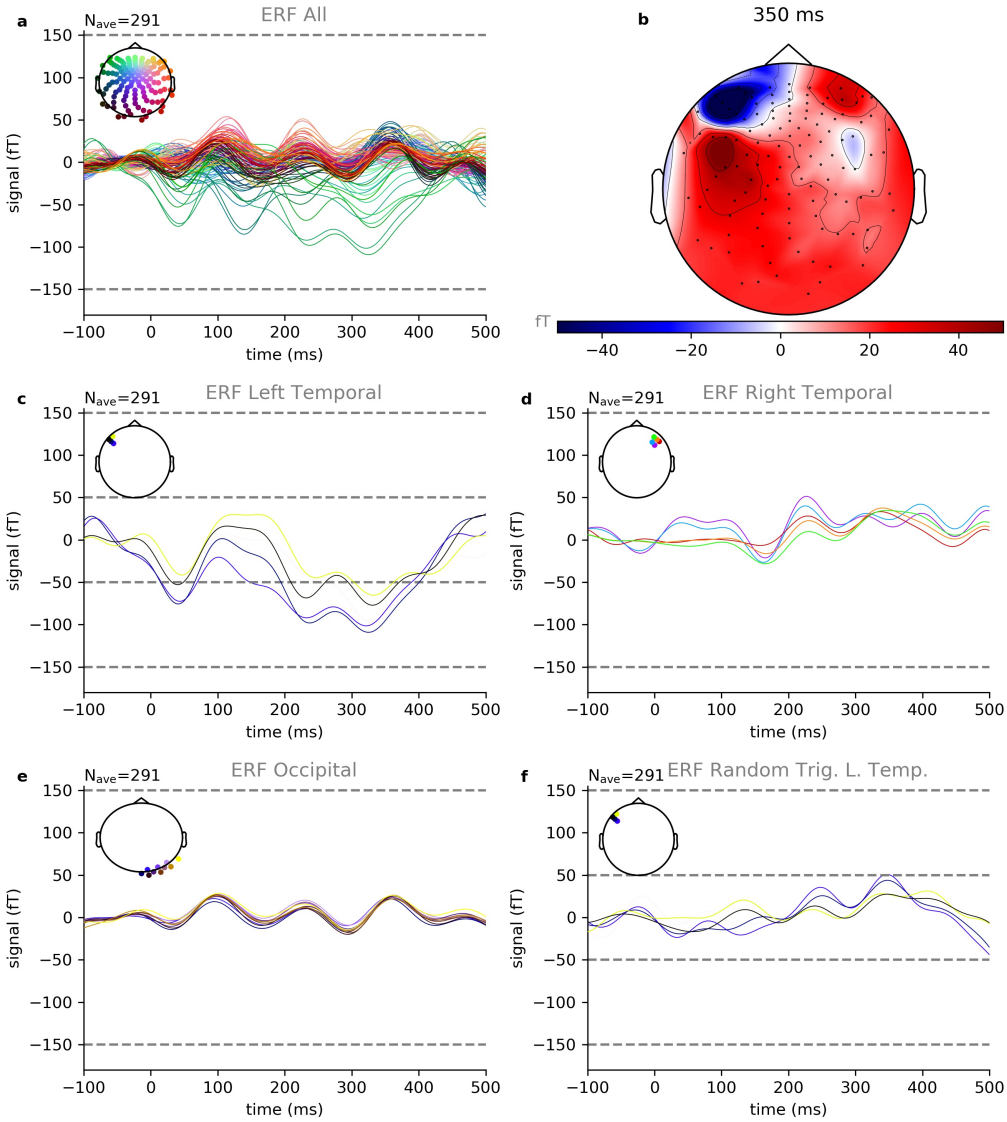

Figure 6: **Event related fields for word onset.** Shown are exemplary data of book parts number 1-3 of 10 from subject 2 of 15. a: Summary of ERFs of all 248 recording channels averaged over 291 trials. Colors indicate the positions of the recording channels. b: Spatial distribution of ERF amplitudes at 350ms after word onset. c: The largest amplitudes occur in channels located at temporal and frontal areas of the left hemisphere. d: The corresponding channels at the right hemisphere show clearly smaller ERF amplitudes. e: The same is true for occipital channels. f: Same channels as in c, but averaged over randomly chosen triggers instead of word onset triggers. Also in this control condition, the resulting amplitudes are smaller than those for the word onset condition.

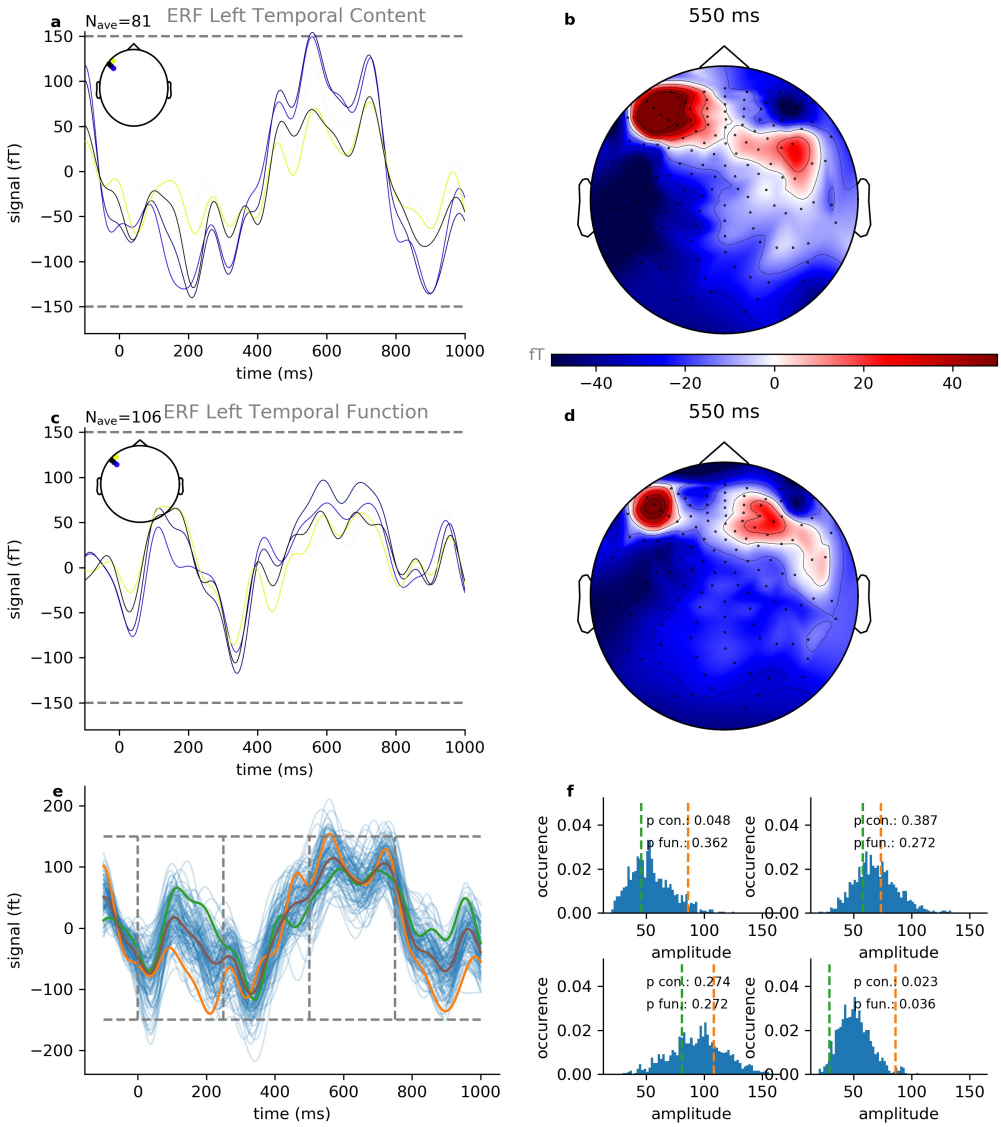

Figure 7: **Event related fields for function and content words.** Shown are exemplary data of book parts number 1-3 of 10 from subject 2 of 15. a: Averaged ERFs for content words (n=81 trials) with largest amplitudes. b: Spatial distribution of ERF amplitudes at 550ms after word onset for content words. c: Averaged ERFs for function words (n=106 trials) with largest amplitudes. d: Spatial distribution of ERF amplitudes at 550ms after word onset for function words. e: ERF with the largest amplitude for content words (orange) and function words (green), together with ERFs derived from permutation test (blue). f: Distribution of ERF amplitudes derived from permutation test within four subsequent time frames: $0\,ms$–$250\,ms$ (upper left), $250\,ms$–$500\,ms$ (upper right), $500\,ms$–$750\,ms$ (lower left) and $750\,ms$–$1000\,ms$ (lower right).

References

[1] Annette MB De Groot and Peter Hagoort. *Research methods in psycholinguistics and the neurobiology of language: A practical guide*, volume 9. John Wiley & Sons, 2017.

[2] Matti Hämäläinen, Riitta Hari, Risto J Ilmoniemi, Jukka Knuutila, and Olli V Lounasmaa. Magnetoencephalography—theory, instrumentation, and applications to noninvasive studies of the working human brain. *Reviews of modern Physics*, 65(2):413, 1993.

[3] David Millett. Hans berger: From psychic energy to the eeg. *Perspectives in biology and medicine*, 44(4):522–542, 2001.

[4] Benjamin Files. An introduction to eeg. *Perception*, 2011.

[5] Friedemann Pulvermüller and Yury Shtyrov. Spatiotemporal signatures of large-scale synfire chains for speech processing as revealed by meg. *Cerebral cortex*, 19(1):79–88, 2008.

[6] Friedemann Pulvermüller, Yury Shtyrov, and Olaf Hauk. Understanding in an instant: neurophysiological evidence for mechanistic language circuits in the brain. *Brain and language*, 110(2):81–94, 2009.

[7] Gwenda L Schmidt-Snoek, Ashley R Drew, Elizabeth C Barile, and Stephen J Agauas. Auditory and motion metaphors have different scalp distributions: an erp study. *Frontiers in human neuroscience*, 9:126, 2015.

[8] Valentina Bambini, Chiara Bertini, Walter Schaeken, Alessandra Stella, and Francesco Di Russo. Disentangling metaphor from context: an erp study. *Frontiers in psychology*, 7:559, 2016.

[9] Vicky T Lai, Olivia Howerton, and Rutvik H Desai. Concrete processing of action metaphors: Evidence from erp. *Brain research*, 1714:202–209, 2019.

[10] Rosario Tomasello, Cora Kim, Felix R Dreyer, Luigi Grisoni, and Friedemann Pulvermüller. Neurophysiological evidence for rapid processing of verbal and gestural information in understanding communicative actions. *Scientific reports*, 9(1):1–17, 2019.

[11] Friedemann Pulvermüller, Hubert Preißl, Carsten Eulitz, Christo Pantev, Werner Lutzenberger, Bernd Feige, Thomas Elbert, and Niels Birbaumer. Gamma-band responses reflect word/pseudoword processing. In *Oscillatory event-related brain dynamics*, pages 243–258. Springer, 1994.

[12] Matt Craddock, Jasna Martinovic, and Matthias M Müller. Early and late effects of objecthood and spatial frequency on event-related potentials and gamma band activity. *BMC neuroscience*, 16(1):6, 2015.

[13] Rachel L Moseley, Friedemann Pulvermüller, and Yury Shtyrov. Sensorimotor semantics on the spot: brain activity dissociates between conceptual categories within 150 ms. *Scientific reports*, 3:1928, 2013.

[14] Angela D Friederici, Christian J Fiebach, Matthias Schlesewsky, Ina D Bornkessel, and D Yves Von Cramon. Processing linguistic complexity and grammaticality in the left frontal cortex. *Cerebral Cortex*, 16(12):1709–1717, 2006.

[15] Nai Ding and Jonathan Z Simon. Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. *Journal of neurophysiology*, 107(1):78–89, 2012.

[16] Lauren J Silbert, Christopher J Honey, Erez Simony, David Poeppel, and Uri Hasson. Coupled neural systems underlie the production and comprehension of naturalistic narrative speech. *Proceedings of the National Academy of Sciences*, 111(43):E4687–E4696, 2014.

[17] Michael P Broderick, Andrew J Anderson, Giovanni M Di Liberto, Michael J Crosse, and Edmund C Lalor. Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech. *Current Biology*, 28(5):803–809, 2018.

[18] Christian Brodbeck, Alessandro Presacco, and Jonathan Z Simon. Neural source dynamics of brain responses to continuous stimuli: Speech processing from acoustics to comprehension. *NeuroImage*, 172:162–174, 2018.

[19] Phillip M Alday. M/eeg analysis of naturalistic stories: a review from speech to language processing. *Language, Cognition and Neuroscience*, 34(4):457–473, 2019.

[20] Fatma Deniz, Anwar O Nunez-Elizalde, Alexander G Huth, and Jack L Gallant. The representation of semantic information across human cerebral cortex during listening versus reading is invariant to stimulus modality. *Journal of Neuroscience*, 39(39):7722–7736, 2019.

[21] Michael GH Coles and Michael D Rugg. *Event-related brain potentials: An introduction.* Oxford University Press, 1995.

[22] Gert Pfurtscheller and FH Lopes Da Silva. Event-related eeg/meg synchronization and desynchronization: basic principles. *Clinical neurophysiology*, 110(11):1842–1857, 1999.

[23] Todd C Handy. *Event-related potentials: A methods handbook*. MIT press, 2005.

[24] Geoffrey F Woodman. A brief introduction to the use of event-related potentials in studies of perception and attention. *Attention, Perception, & Psychophysics*, 72(8):2031–2046, 2010.

[25] SJ Luck. An introduction to the event-related potential technique, second, 2014.

[26] Kalanit Grill-Spector, Richard Henson, and Alex Martin. Repetition and the brain: neural models of stimulus-specific effects. *Trends in cognitive sciences*, 10(1):14–23, 2006.

[27] Sarah Laszlo and Kara D Federmeier. The n400 as a snapshot of interactive processing: Evidence from regression analyses of orthographic neighbor and lexical associate effects. *Psychophysiology*, 48(2):176–186, 2011.

[28] Leila Wehbe, Brian Murphy, Partha Talukdar, Alona Fyshe, Aaditya Ramdas, and Tom Mitchell. Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses. *PloS one*, 9(11):e112575, 2014.

[29] Alexander G Huth, Wendy A De Heer, Thomas L Griffiths, Frédéric E Theunissen, and Jack L Gallant. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532(7600):453, 2016.

[30] Clive Souter and Eric Atwell. *Corpus-based computational linguistics*. Number 9. Rodopi, 1993.

[31] John Sinclair. Trust the text. In *Trust the text*, pages 19–33. Routledge, 2004.

[32] Guy Aston and Lou Burnard. *The BNC handbook: exploring the British National Corpus with SARA*. Capstone, 1998.

[33] Nancy Ide and Keith Suderman. The american national corpus first release. In *LREC*, 2004.

[34] Mark Davies. The corpus of contemporary american english as the first reliable monitor corpus of english. *Literary and linguistic computing*, 25(4):447–464, 2010.

[35] Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K Gray, Joseph P Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, et al. Quantitative analysis of culture using millions of digitized books. *science*, 331(6014):176–182, 2011.

[36] Dennis A Trinkle, Dorothy Auchter, Scott A Merriman, and Todd E Larson. *The History Highway: A 21st-century Guide to Internet Resources*. Routledge, 2016.

[37] Roland Schäfer and Felix Bildhauer. Building large corpora from the web using a new efficient tool chain. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 486–493, Istanbul, Turkey, 2012. European Language Resources Association (ELRA).

[38] Stefan Evert. *The statistics of word cooccurrences: word pairs and collocations*. 2005.

[39] M Koskinen and Mika Seppä. Uncovering cortical meg responses to listened audiobook stories. *Neuroimage*, 100:263–270, 2014.

[40] Geoffrey Leech. The state of the art in corpus linguistics. In *English corpus linguistics*, pages 20–41. Routledge, 2014.

[41] Léon Bottou, Corinna Cortes, John S Denker, Harris Drucker, Isabelle Guyon, Larry D Jackel, Yann LeCun, Urs A Muller, Edward Sackinger, Patrice Simard, et al. Comparison of classifier methods: a case study in handwritten digit recognition. In *Proceedings of the 12th IAPR International Conference on Pattern Recognition, Vol. 3-Conference C: Signal Processing (Cat. No. 94CH3440-5)*, volume 2, pages 77–82. IEEE, 1994.

[42] Norbert L Kerr. Harking: Hypothesizing after the results are known. *Personality and Social Psychology Review*, 2(3):196–217, 1998.

[43] Marcus R Munafò, Brian A Nosek, Dorothy VM Bishop, Katherine S Button, Christopher D Chambers, Nathalie Percie Du Sert, Uri Simonsohn, Eric-Jan Wagenmakers, Jennifer J Ware, and John PA Ioannidis. A manifesto for reproducible science. *Nature human behaviour*, 1(1):0021, 2017.

[44] Friedemann Pulvermüller, Werner Lutzenberger, and Niels Birbaumer. Electrocortical distinction of vocabulary types. *Electroencephalography and clinical Neurophysiology*, 94(5):357–370, 1995.

[45] Michele T Diaz and Gregory McCarthy. A comparison of brain activity evoked by single content and function words: an fmri investigation of implicit word processing. *Brain research*, 1282:38–49, 2009.

[46] Richard C Oldfield. The assessment and analysis of handedness: the edinburgh inventory. *Neuropsychologia*, 9(1):97–113, 1971.

[47] Adwait Ratnaparkhi. A maximum entropy model for part-of-speech tagging. In *Conference on Empirical Methods in Natural Language Processing*, 1996.

[48] Lluís Màrquez and Horacio Rodríguez. Part-of-speech tagging using decision trees. In *European Conference on Machine Learning*, pages 25–36. Springer, 1998.

[49] AI Explosion. spacy-industrial-strength natural language processing in python. *URL: https://spacy. io*, 2017.

[50] Pedro J Moreno, Chris Joerg, Jean-Manuel Van Thong, and Oren Glickman. A recursive algorithm for the forced alignment of very long audio

segments. In *Fifth International Conference on Spoken Language Processing*, 1998.

[51] Jiahong Yuan and Mark Liberman. Investigating/l/variation in english through forced alignment. In *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[52] Athanasios Katsamanis, Matthew Black, Panayiotis G Georgiou, Louis Goldstein, and S Narayanan. Sailalign: Robust long speech-text alignment. In *Proc. of Workshop on New Tools and Methods for Very-Large Scale Phonetics Research*, 2011.

[53] Florian Schiel. Automatic Phonetic Transcription of Non-Prompted Speech. In *Proc. of the ICPhS*, pages 607–610, San Francisco, August 1999.

[54] Thomas Kisler, Uwe Reichel, and Florian Schiel. Multilingual processing of speech via web services. *Computer Speech & Language*, 45:326–347, 2017.

[55] Jonathan W Peirce. Psychopy—psychophysics software in python. *Journal of neuroscience methods*, 162(1-2):8–13, 2007.

[56] Jonathan W Peirce. Generating stimuli for neuroscience using psychopy. *Frontiers in neuroinformatics*, 2:10, 2009.

[57] Alexandre Gramfort, Martin Luessi, Eric Larson, Denis A Engemann, Daniel Strohmeier, Christian Brodbeck, Roman Goj, Mainak Jas, Teon Brooks, Lauri Parkkonen, et al. Meg and eeg data analysis with mne-python. *Frontiers in neuroscience*, 7:267, 2013.

[58] Alexandre Gramfort, Martin Luessi, Eric Larson, Denis A Engemann, Daniel Strohmeier, Christian Brodbeck, Lauri Parkkonen, and Matti S Hämäläinen. Mne software for processing meg and eeg data. *Neuroimage*, 86:446–460, 2014.

[59] Manfred Fuchs, Jörn Kastner, Michael Wagner, Susan Hawes, and John S Ebersole. A standardized boundary element method volume conductor model. *Clinical Neurophysiology*, 113(5):702–712, 2002.

[60] Alan C Evans, Andrew L Janke, D Louis Collins, and Sylvain Baillet. Brain templates and atlases. *Neuroimage*, 62(2):911–922, 2012.

[61] Olaf Hauk, Matthew H Davis, M Ford, Friedemann Pulvermüller, and William D Marslen-Wilson. The time course of visual word recognition as revealed by linear regression analysis of erp data. *Neuroimage*, 30(4):1383–1400, 2006.

[62] Yury Shtyrov and Friedemann Pulvermüller. Early meg activation dynamics in the left temporal and inferior frontal cortex reflect semantic context integration. *Journal of Cognitive Neuroscience*, 19(10):1633–1642, 2007.

[63] Andreas Højlund, Line Gebauer, William B McGregor, and Mikkel Wallentin. Context and perceptual asymmetry effects on the mismatch negativity (mmnm) to speech sounds: an meg study. *Language, Cognition and Neuroscience*, 34(5):545–560, 2019.

[64] Angela D Friederici and Sarah ME Gierhan. The language network. *Current opinion in neurobiology*, 23(2):250–254, 2013.

[65] Marta Kutas and Kara D Federmeier. Thirty years and counting: finding meaning in the n400 component of the event-related brain potential (erp). *Annual review of psychology*, 62:621–647, 2011.

[66] Angela D Friederici, Erdmut Pfeifer, and Anja Hahne. Event-related brain potentials during natural speech processing: Effects of semantic, morphological and syntactic violations. *Cognitive brain research*, 1(3):183–192, 1993.

[67] Antje Strauß, Sonja A Kotz, and Jonas Obleser. Narrowed expectancies under degraded speech: revisiting the n400. *Journal of Cognitive Neuroscience*, 25(8):1383–1395, 2013.

[68] Peter Hagoort and Colin M Brown. Erp effects of listening to speech: semantic erp effects. *Neuropsychologia*, 38(11):1518–1530, 2000.

[69] Ellen Lau, Diogo Almeida, Paul C Hines, and David Poeppel. A lexical basis for n400 context effects: Evidence from meg. *Brain and language*, 111(3):161–172, 2009.

[70] Friedemann Pulvermüller. *The neuroscience of language: On brain circuits of words and serial order*. Cambridge University Press, 2003.

[71] David Kemmerer. *Cognitive neuroscience of language*. Psychology Press, 2014.

[72] Jonathan Brennan and Liina Pylkkänen. The time-course and spatial distribution of brain activity associated with sentence processing. *Neuroimage*, 60(2):1139–1148, 2012.

[73] Jonathan R Brennan and Liina Pylkkänen. Meg evidence for incremental sentence composition in the anterior temporal lobe. *Cognitive Science*, 41:1515–1531, 2017.

[74] Mario E Archila-Meléndez, Vivianne H Kranen-Mastenbroek, Giancarlo Valente, João Correia, Erik D Gommer, Bernadette M Jansma, Rob P Rouhl, and Mark J Roberts. S09. the role of oscillatory activity in attentive speech perception: An ecog study in epilepsy patients. *Clinical Neurophysiology*, 129:e145, 2018.

[75] Nathan Weisz, Thomas Hartmann, Nadia Müller, and Jonas Obleser. Alpha rhythms in audition: cognitive and clinical perspectives. *Frontiers in psychology*, 2:73, 2011.

[76] Nadia Müller and Nathan Weisz. Lateralized auditory cortical alpha band activity and interregional connectivity pattern reflect anticipation of target sounds. *Cerebral Cortex*, 22(7):1604–1613, 2012.

[77] Marcel Bastiaansen and Peter Hagoort. Oscillatory neuronal dynamics during language comprehension. *Progress in brain research*, 159:179–196, 2006.

[78] Edith Kaan and Tamara Y Swaab. The brain circuitry of syntactic comprehension. *Trends in cognitive sciences*, 6(8):350–356, 2002.

[79] Elizabeth Bates. Processing complex sentences: A cross-linguistic study. *Language and cognitive processes*, 14(1):69–123, 1999.

[80] Stefan Evert, Peter Uhrig, Sabine Bartsch, and Thomas Proisl. E-VIEW-alation – a large-scale evaluation study of association measures for collocation identification. In Iztok Kosem, Carole Tiberius, Miloš

Jakubíček, Jelena Kallas, Simon Krek, and Vít Baisa, editors, *Electronic lexicography in the 21st century. Proceedings of the eLex 2017 conference*, pages 531–549, Leiden, 2017. Lexical Computing.

[81] Adele E Goldberg. *Constructions: A construction grammar approach to argument structure*. University of Chicago Press, 1995.

[82] Adele E Goldberg. Constructions: A new theoretical approach to language. *Trends in cognitive sciences*, 7(5):219–224, 2003.

[83] Adele E Goldberg. *Constructions at work: The nature of generalization in language*. Oxford University Press on Demand, 2006.

[84] Thomas Herbst and Susen Schüller. *Introduction to syntactic analysis: A valency approach*. Narr Francke Attempto Verlag, 2008.

[85] Thomas Herbst. The status of generalizations: Valency and argument structure constructions. *Zeitschrift für Anglistik und Amerikanistik*, 59(4):347–368, 2011.

[86] Thomas Herbst. The valency approach to argument structure constructions. *Constructions–collocations–patterns*, pages 167–216, 2014.

[87] Şeyda Özçalışkan and Susan Goldin-Meadow. Sex differences in language first appear in gesture. *Developmental science*, 13(5):752–760, 2010.

[88] Alice Mado Proverbio, Marta Calbi, Mirella Manfredi, and Alberto Zani. Comprehending body language and mimics: an erp and neuroimaging study on italian actors and viewers. *PLoS One*, 9(3), 2014.

[89] Patrick Krauss, Claus Metzner, Achim Schilling, Konstantin Tziridis, Maximilian Traxdorf, Andreas Wollbrink, Stefan Rampp, Christo Pantev, and Holger Schulze. A statistical method for analyzing and comparing spatiotemporal cortical activation patterns. *Scientific reports*, 8(1):5433, 2018.

[90] Achim Schilling, Claus Metzner, Jonas Rietsch, Richard Gerum, Holger Schulze, and Patrick Krauss. How deep is deep enough?–quantifying class separability in the hidden layers of deep neural networks. *arXiv preprint arXiv:1811.01753*, 2018.

[91] Achim Schilling, Richard Carl Gerum, Patrick Krauss, Claus Metzner, Konstantin Tziridis, and Holger Schulze. Objective estimation of sensory thresholds based on neurophysiological parameters. *Frontiers in neuroscience*, 13:481, 2019.

[92] Peter Uhrig. I don't want to go all yoko ono on you – creativity and variation in a family of constructions. *Zeitschrift für Anglistik und Amerikanistik*, 66(3):295–308, 2018.

[93] Peter Uhrig. Creative intentions: The thin line between 'creative' and 'wrong'. *Cognitive Semiotics*, 13(1), 2020.

[94] Nikolaus Kriegeskorte and Pamela K Douglas. Cognitive computational neuroscience. *Nature neuroscience*, 21(9):1148–1160, 2018.

Additional Information

Acknowledgments

This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation): grant KR5148/2-1 to PK – project number 436456810, the Interdisciplinary Center for Clinical Research (IZKF) at the University Hospital of the University Erlangen-Nuremberg (grant ELAN-17-12-27-1-Schilling to AS), and the Emergent Talents Initiative (ETI) of the University Erlangen-Nuremberg (grant 2019/2-Phil-01 to PK).
We are grateful to the publishers *Deuticke Verlag* and *Hörbuch Hamburg* for the permission to use the novel and corresponding audiobook *Gut gegen Nordwind* by Daniel Glattauer for the present and future studies.
We thank Martin Kaltenhäuser for technical assistance and Stefan Rampp for useful discussion.

Competing interests

The authors declare no competing financial interests.

Author contributions

PK and AS designed the study. AS, PK and VK prepared the stimulation and processed the audiobook. VK, MH, AS and PK performed the experiments. AS and PK analyzed the data. AS, PK and KS analyzed the text of the novel. AS, PK, RT, MRHS, AM and PU discussed the results. PK, AS, RT, MRHS and PU wrote the manuscript.

Supplements

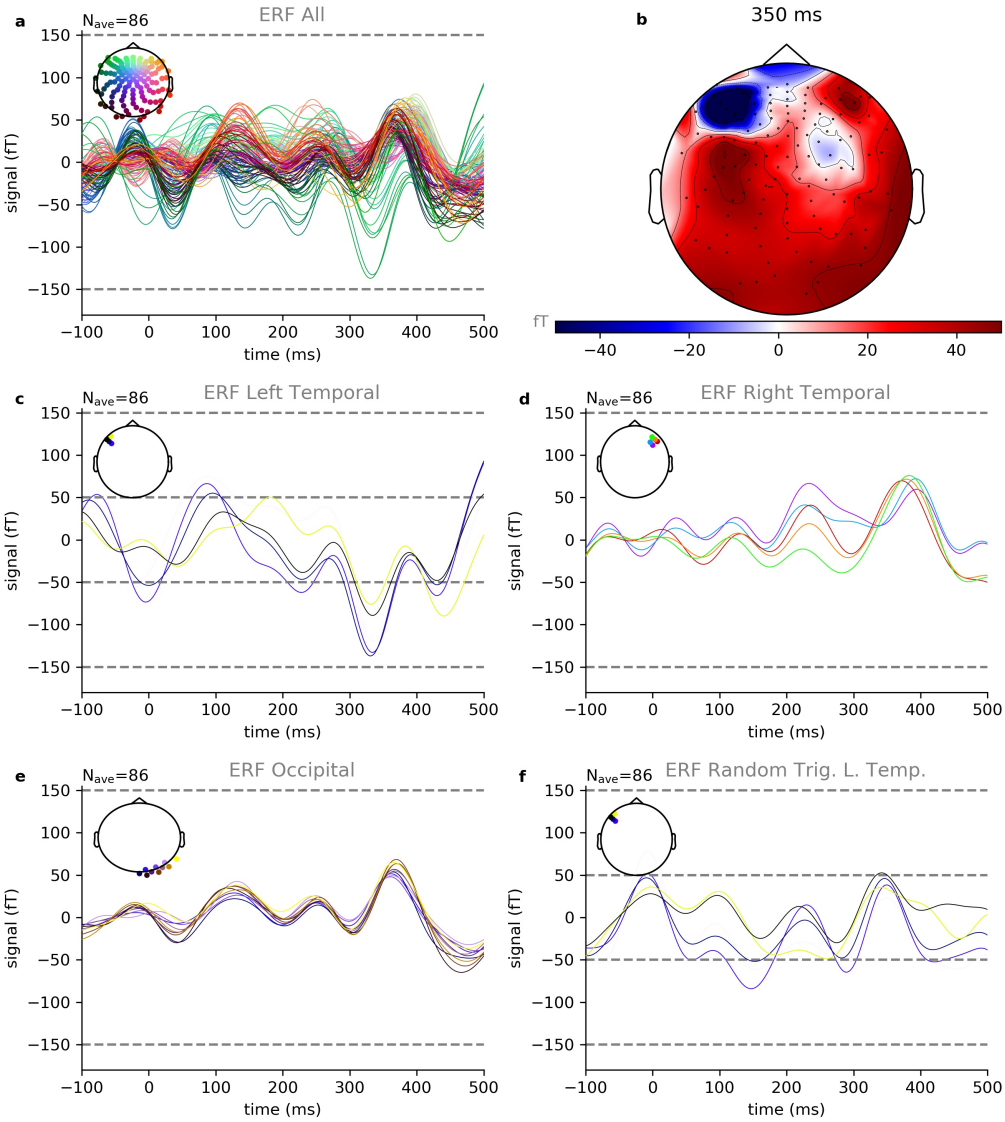

Figure S1: **Event related fields of word onsets. Sub-sample 1 of 3.** Shown are exemplary data of book part number 1 of 10 from subject 2 of 15. a: Summary of ERFs of all 248 recording channels averaged over 86 trials. Colors indicate the positions of the recording channels. b: Spatial distribution of ERF amplitudes at 350ms after word onset. c: The largest amplitudes occur in channels located at temporal and frontal areas of the left hemisphere. d: The corresponding channels at the right hemisphere show clearly smaller ERF amplitudes. e: The same is true for occipital channels. f: Same channels as in c, but averaged over randomly chosen triggers instead of word onset triggers. Also in this control condition, the resulting amplitudes are smaller than those for the word onset condition.

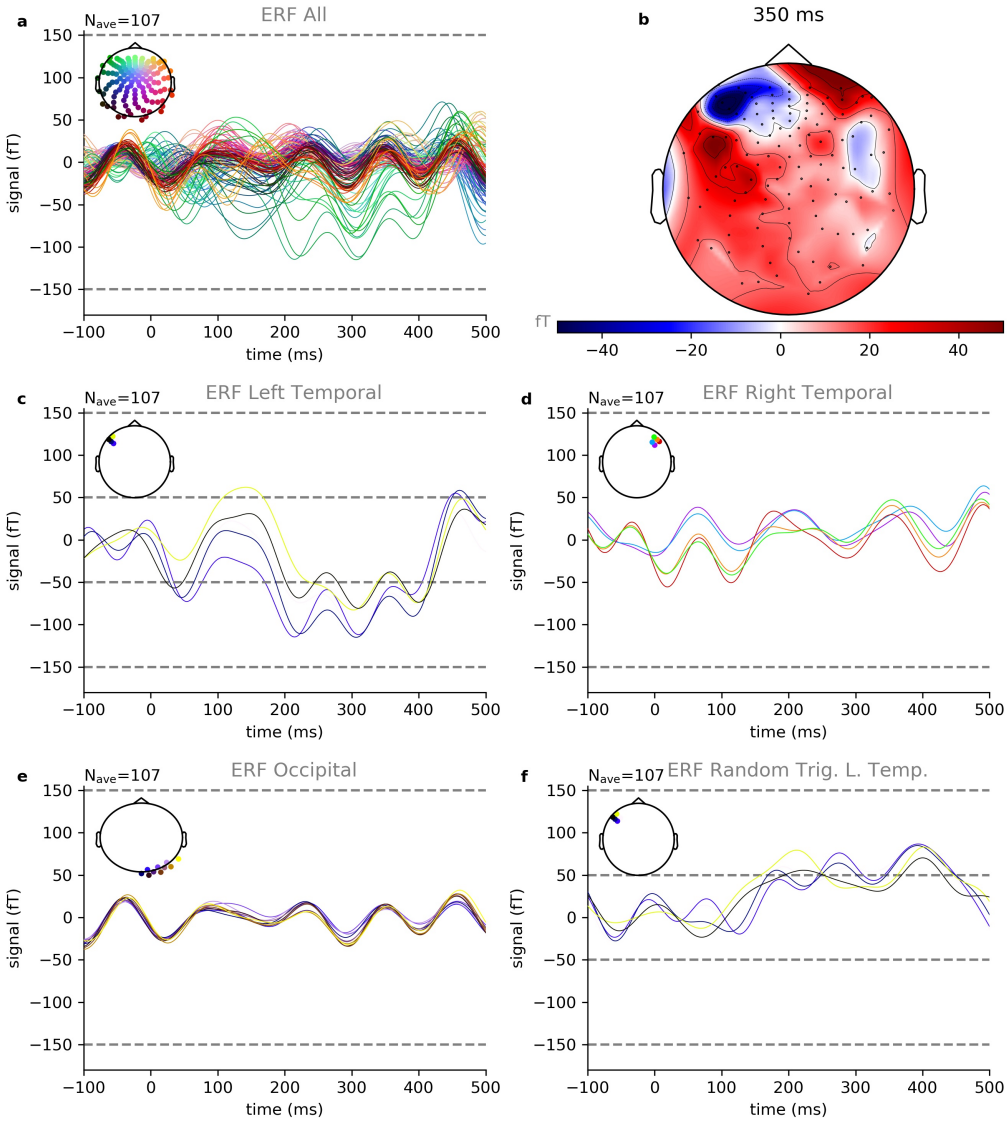

Figure S2: **Event related fields of word onsets. Sub-sample 2 of 3.** Shown are exemplary data of book part number 2 of 10 from subject 2 of 15. a: Summary of ERFs of all 248 recording channels averaged over 107 trials. Colors indicate the positions of the recording channels. b: Spatial distribution of ERF amplitudes at 350ms after word onset. c: The largest amplitudes occur in channels located at temporal and frontal areas of the left hemisphere. d: The corresponding channels at the right hemisphere show clearly smaller ERF amplitudes. e: The same is true for occipital channels. f: Same channels as in c, but averaged over randomly chosen triggers instead of word onset triggers. Also in this control condition, the resulting amplitudes are smaller than those for the word onset condition.

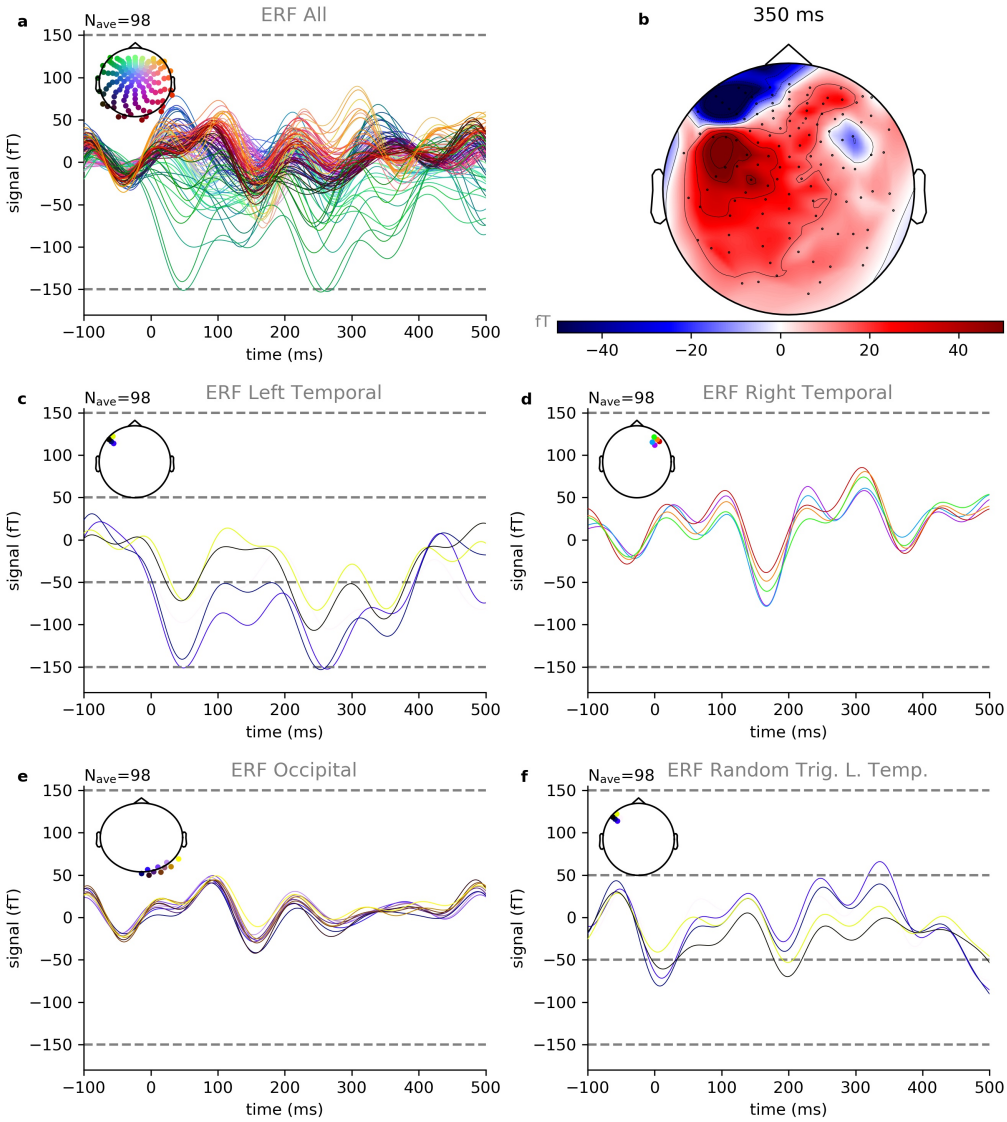

Figure S3: **Event related fields of word onsets. Sub-sample 3 of 3.** Shown are exemplary data of book part number 3 of 10 from subject 2 of 15. a: Summary of ERFs of all 248 recording channels averaged over 98 trials. Colors indicate the positions of the recording channels. b: Spatial distribution of ERF amplitudes at 350ms after word onset. c: The largest amplitudes occur in channels located at temporal and frontal areas of the left hemisphere. d: The corresponding channels at the right hemisphere show clearly smaller ERF amplitudes. e: The same is true for occipital channels. f: Same channels as in c, but averaged over randomly chosen triggers instead of word onset triggers. Also in this control condition, the resulting amplitudes are smaller than those for the word onset condition.

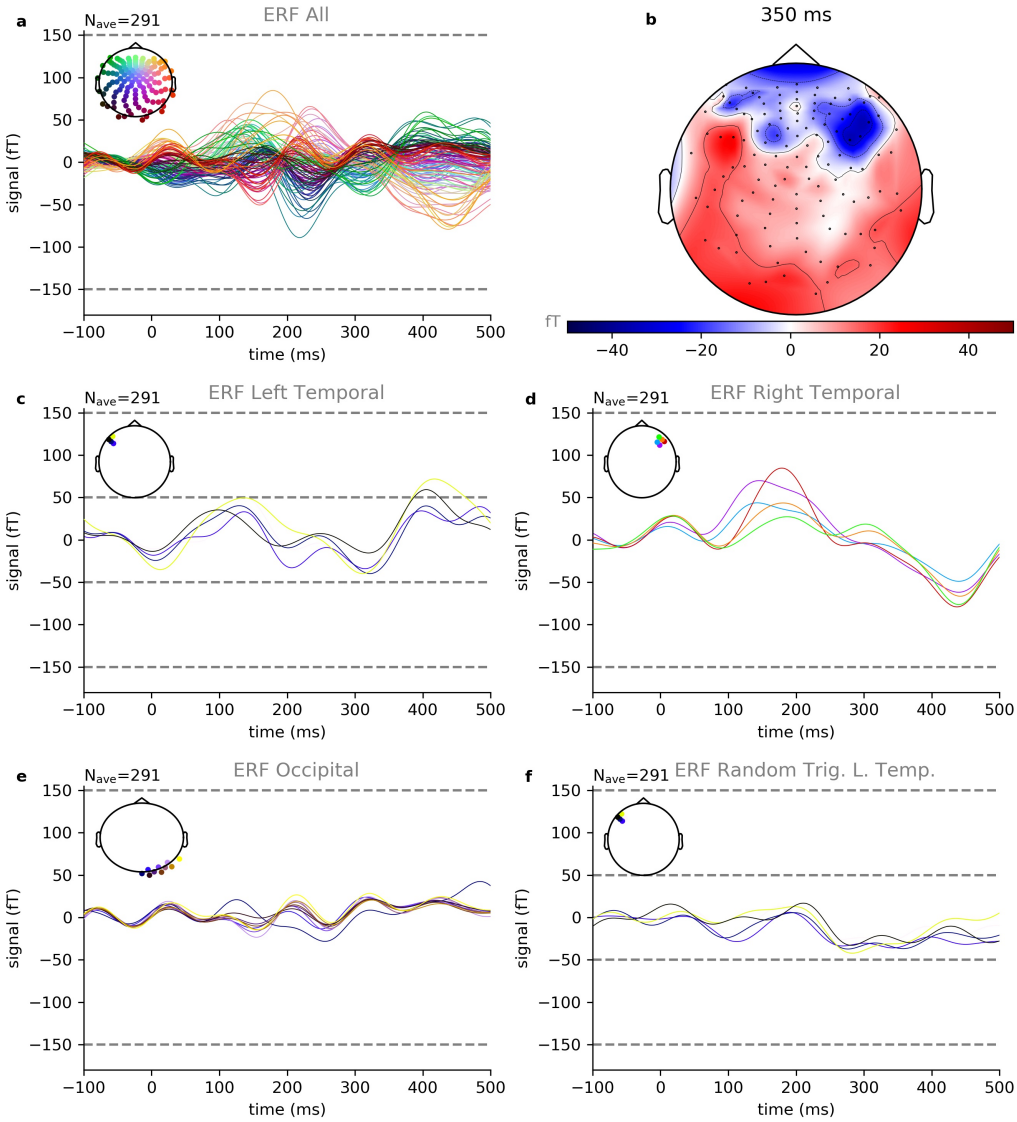

Figure S4: **Event related fields for word onset.** Shown are exemplary data of book parts number 1-3 of 10 from subject 5 of 15. a: Summary of ERFs of all 248 recording channels averaged over 291 trials. Colors indicate the positions of the recording channels. b: Spatial distribution of ERF amplitudes at 350ms after word onset. c: The largest amplitudes occur in channels located at temporal and frontal areas of the left hemisphere. d: The corresponding channels at the right hemisphere show clearly smaller ERF amplitudes. e: The same is true for occipital channels. f: Same channels as in c, but averaged over randomly chosen triggers instead of word onset triggers. Also in this control condition, the resulting amplitudes are smaller than those for the word onset condition.

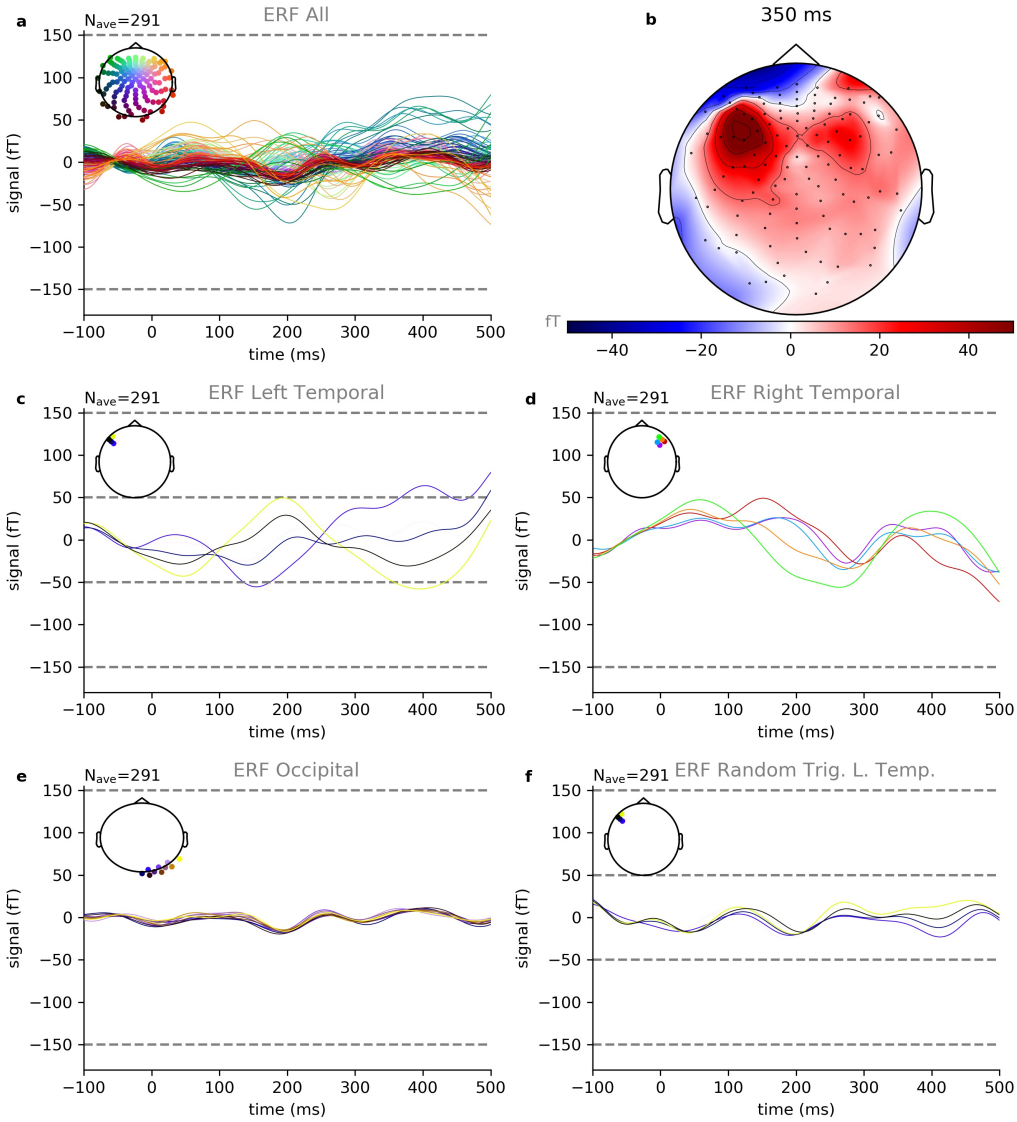

Figure S5: **Event related fields for word onset.** Shown are exemplary data of book parts number 1-3 of 10 from subject 8 of 15. a: Summary of ERFs of all 248 recording channels averaged over 291 trials. Colors indicate the positions of the recording channels. b: Spatial distribution of ERF amplitudes at 350ms after word onset. c: The largest amplitudes occur in channels located at temporal and frontal areas of the left hemisphere. d: The corresponding channels at the right hemisphere show clearly smaller ERF amplitudes. e: The same is true for occipital channels. f: Same channels as in c, but averaged over randomly chosen triggers instead of word onset triggers. Also in this control condition, the resulting amplitudes are smaller than those for the word onset condition.

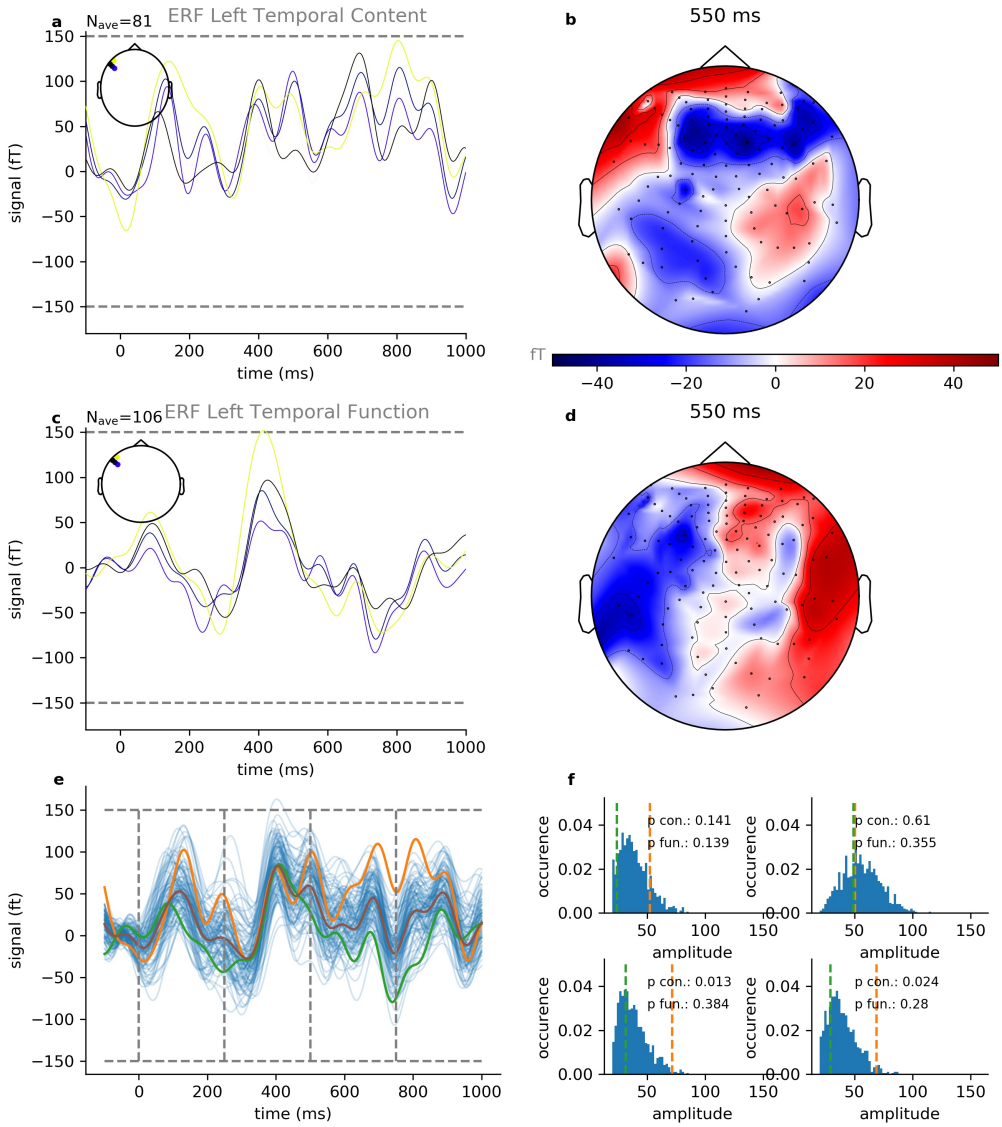

Figure S6: **Event related fields for function and content words.** Shown are exemplary data of book parts number 1-3 of 10 from subject 5 of 15. a: Averaged ERFs for content words (n=81 trials) with largest amplitudes. b: Spatial distribution of ERF amplitudes at 550ms after word onset for content words. c: Averaged ERFs for function words (n=106 trials) with largest amplitudes. d: Spatial distribution of ERF amplitudes at 550ms after word onset for function words. e: ERF with the largest amplitude for content words (orange) and function words (green), together with ERFs derived from permutation test (blue). f: Distribution of ERF amplitudes derived from permutation test within four subsequent time frames: $0\,ms$–$250\,ms$ (upper left), $250\,ms$–$500\,ms$ (upper right), $500\,ms$–$750\,ms$ (lower left) and $750\,ms$–$1000\,ms$ (lower right).

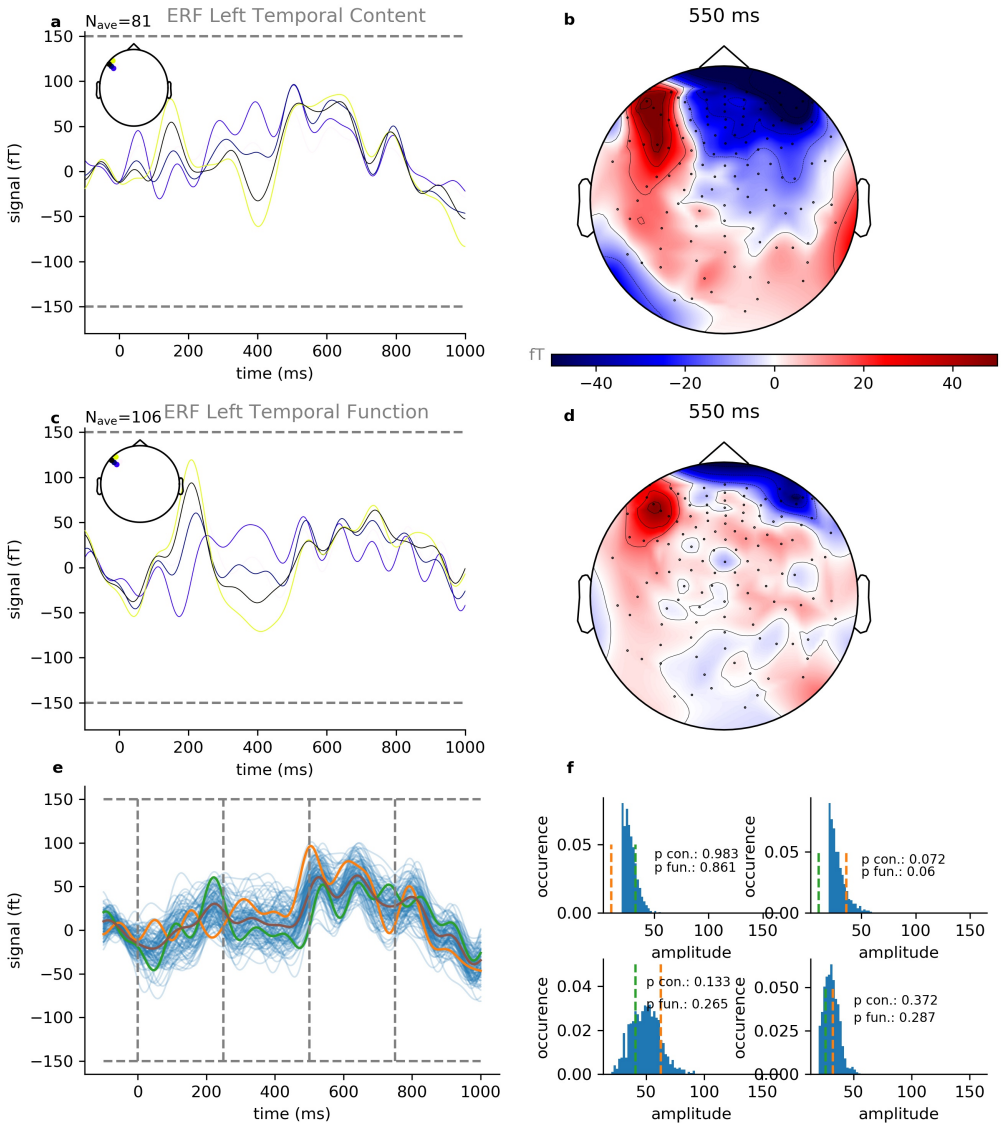

Figure S7: **Event related fields for function and content words.** Shown are exemplary data of book parts number 1-3 of 10 from subject 8 of 15. a: Averaged ERFs for content words (n=81 trials) with largest amplitudes. b: Spatial distribution of ERF amplitudes at 550ms after word onset for content words. c: Averaged ERFs for function words (n=106 trials) with largest amplitudes. d: Spatial distribution of ERF amplitudes at 550ms after word onset for function words. e: ERF with the largest amplitude for content words (orange) and function words (green), together with ERFs derived from permutation test (blue). f: Distribution of ERF amplitudes derived from permutation test within four subsequent time frames: $0\,ms$–$250\,ms$ (upper left), $250\,ms$–$500\,ms$ (upper right), $500\,ms$–$750\,ms$ (lower left) and $750\,ms$–$1000\,ms$ (lower right).

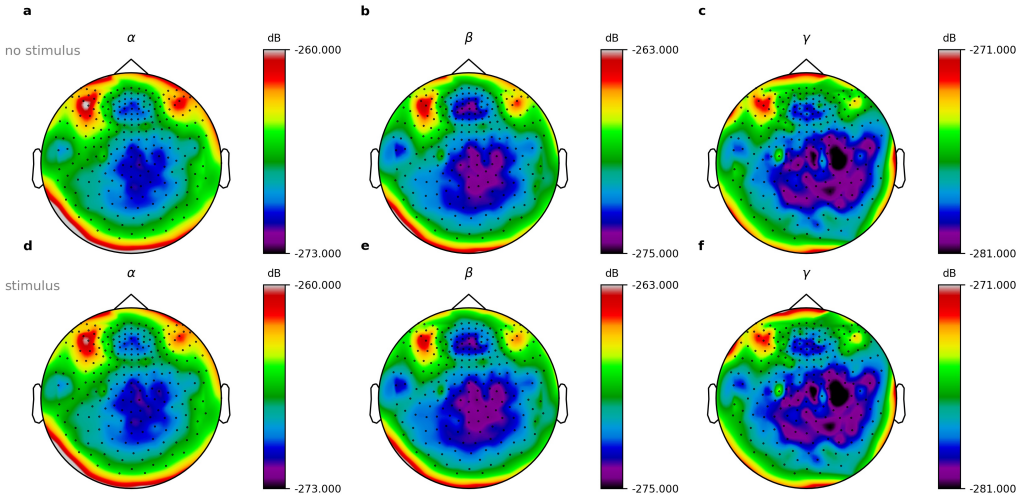

Figure S8: **Normalized power spectra for words and silence.** Shown are exemplary data of book parts number 1-3 of 10 from subject 2 of 15. a-c: Power spectra for word offset, i.e. silence. d-f: Power spectra for word onsets. a,d: Alpha frequency range. b,e: Beta frequency range. c,f: Gamma frequency range.

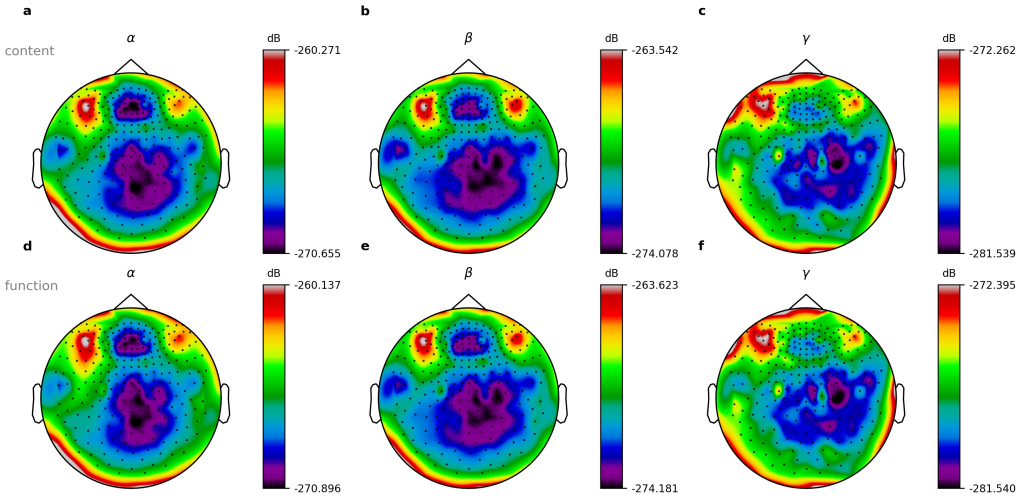

Figure S9: **Normalized power spectra for content and function words.** Shown are exemplary data of book parts number 1-3 of 10 from subject 2 of 15. a-c: Power spectra for content words. d-f: Power spectra for function words. a,d: Alpha frequency range. b,e: Beta frequency range. c,f: Gamma frequency range.

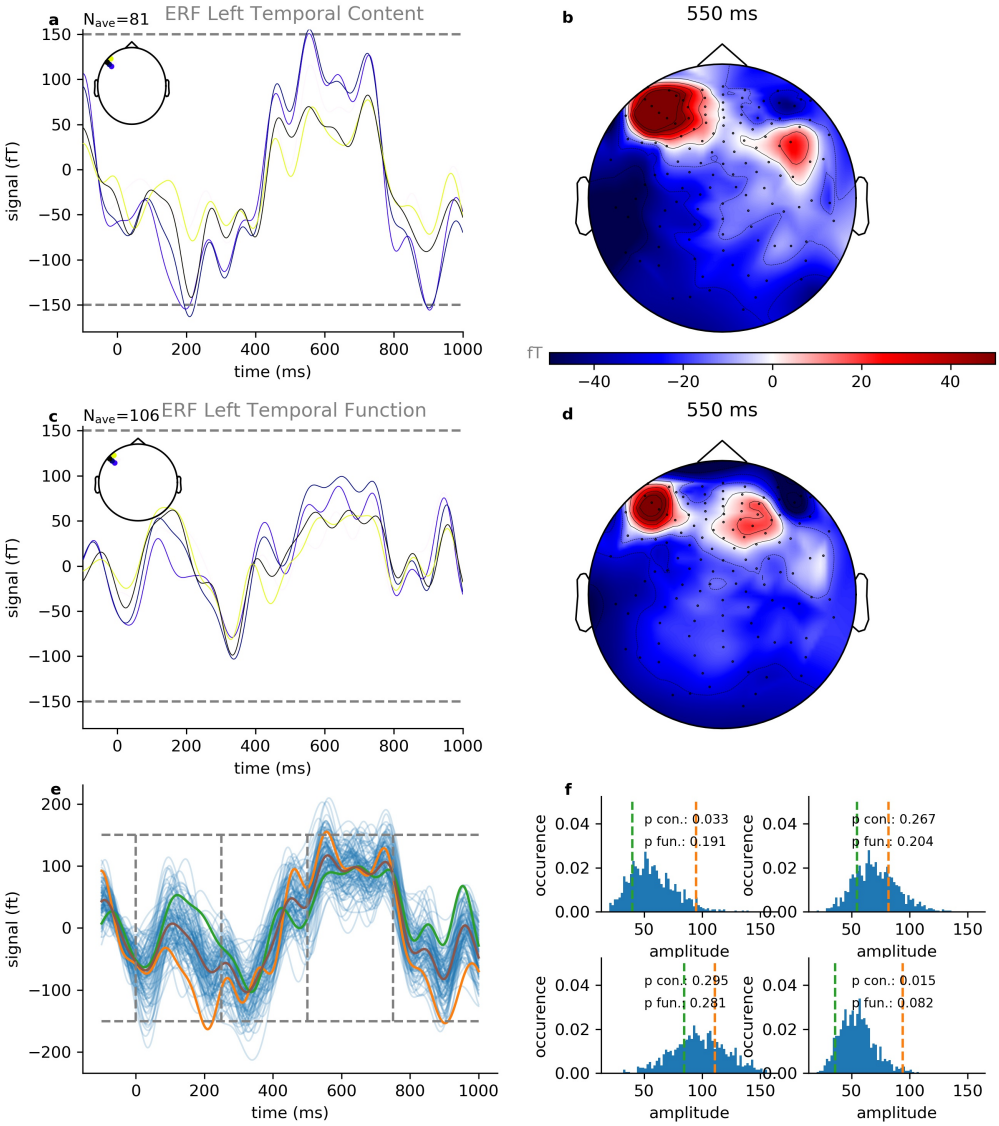

Figure S10: **Event related fields for function and content words.** Shown are the same analyses and participant as in Figure 7. However, for these analyses we did not delete the first two independent components. As the results are qualitatively equal, this does not affect the difference of neural responses to content and function words.